Folksonomies as Document Summary Engines
A Study of *Delicious.com* Tagging Behaviour.

Stephen Wattam, Paul Rayson, Damon Berridge
School of Computing and Communications
University of Lancaster
{*s.wattam, p.rayson, d.berridge*}*@lancaster.ac.uk*

Abstract

Folksonomies like Delicious.com's tagging ecosystem ostensibly offer a ready supply of human-reviewed text and summary information. The manually assigned tags in folksonomies can be seen as the manual counterpart to key words in corpus linguistics that are extracted automatically through tools such as WordSmith (Scott (1996)) and Wmatrix (Rayson (2008)).

This paper examines the Delicious.com social bookmarking service to assess the extent to which tag-based folksonomies of documents represent information already accessible using existing information retrieval metrics. The metrics compared include five different information retrieval measures (Log-Likelihood, TFIDF, RFR, WordNet synonym connectedness and position of first occurrence), as well as two measures derived from tags themselves.

We perform multiple keyword analyses, comparing the results of each against the manual tag selection through dimensionality reduction using principal component analysis. We then go on to fit classifiers to identify tags from keyword selections, and critique the quality of these classifiers' output.

We find that log-likelihood is the single best indicator of manual tag selection, and that improving the classification model by inclusion of other metrics provides only modest improvements in exchange for large increases in complexity.

1 Introduction

With the rise of 'Web 2.0', a dramatic increase in document-folksonomy collections has occurred online. Folksonomies contain, along with the original text, some taxonomic information that may be useful to summarisation and document comparison methods (instead of inferring from small features) (Mathes (2004)).

Taking Delicious.com as a prime example of this pairing, we examine the extent to which tags provide human insight into the document content, and see if this provides a richer information source than simply extracting key words from the document itself. We do this by using conventional salience metrics commonly used for keywording, and by comparing the results to a corpus of tag-document pairs.

This study examines Delicious.com tags against a series of common keywording systems, in order to characterise the relation between document and tag. Because keywords are necessarily extracted from document contents, this naturally leads us to conclusions regarding the content of tagging systems, and the degree to which human judgement and analysis are used when assigning them to documents.

Many folksonomies offer suggested tags, or an auto-tagging service for bulk imports. The second part of this study assesses the use of keyword metrics as a mechanism for providing this function where a large corpus of existing tagged documents does not exist. Taking the conclusions from the exploratory part, we are able to better characterise the tags selected in terms of how they relate to their source document.

Section 2 presents an overview of the folksonomies detailed in this paper and relevant literature used to provide the framework we use to reason about tags taken from Delicious.com.
Section 3 describes the method by which data was processed — how tags were split into terms, and how each of the keywording metrics was calculated from the document collection. The second part of the section discusses the statistical methods used to explore the data.
Section 4 outlines the process of dimensionality reduction, first the mechanical process and fitting, then a discussion of the implications of the results with a view to predicting tags.
Section 5 covers the training of classifiers to predict tags from terms found within documents. A comparison of each of the metrics is refined using the results from the PCA into a more effective, yet more complex, model.
Section 6 discusses the findings of the exploration, and identifies further work regarding tags and the relationship they have to their source documents.

2 Folksonomies

Many subjective observations of folksonomies have been made, largely commenting upon the benefits compared to conventional taxonomies. Fewer comparisons have been made to search systems, to which the data organisation is arguably closer. Although folksonomies' strengths lie in their statistical nature, their lookup techniques and structure are often expressed in taxonomic terms, *i.e.* the use of token-based rather than similarity-based systems.

It has typically been taken for granted that, in the process of creating folksonomies, people inject a degree of human judgement, producing high-level categorisations that go beyond simple content descriptions, however, this is also the goal of a good search algorithm: to group traits of documents using emergent patterns.

Users are presented, in the case of Delicious.com, with a taxonomy-based browsing interface. The effects of this interface were investigated by Guy and Tonkin (2006), who present an overview of tagging behaviour on Delicious.com and provide some useful analyses on compound tagging and taxonomic behaviour. They propose a number of methods to improve one's own tag set for recall using current systems, as well as suggesting that interfaces may be improved beyond the simple tag recall systems they still represent today. Their data on compound tag usage is used to inform our own tag processing and normalisation, allowing comparison to document keywords.

Hassan-Montero and Herrero-Solana (2006) investigate some of these methods, adding a statistical angle on tag recall that mirrors the methods used in language. They present tag versions of TFIDF, based around document counts, as well as a reformed user interface based on the metric. This metric is used as our normalised tag frequency measure, and is compared to conventional TFIDF scores for keywords in sections 3, 4 and 5.

Combining this statistical view of tags with conventional document measures is examined in depth for a single document by Kehoe and Gee (2011), who inspect the mechanisms behind tagging at a fine-grained level by analysing a single document with relation to both its tags and its keywords, as identified using a log-likelihood statistic. Their classification of tags which do not appear in the text itself lends a framework which may be used to extend reasoning beyond the model we present here, possibly providing hints that a computational approach to keyword generation may approximate non-document tags more appropriately.

The current study may be thought of as a more comprehensive extension of Kehoe and Gee (2011)'s inspection of the behaviour of keywording algorithms relative to human tagging. Using observations of tags from Delicious.com we have attempted to solve the main problems of synonymy and syntactic variance through use of common linguistic methods, and have included a

metric that is able to assess the degree to which polysemy and the specificity of terms affects the use of tags.

3 Method

3.1 Data Source

Data were taken from submissions to Delicious.com during September 2009. Since the data were without document contents, these needed downloading and processing into plain text before analyses could be carried out. Documents were downloaded using a series of shell scripts, designed to appear to a server as if they were the Firefox web browser running on Windows XP.

The downloaded documents were then passed through a series of filters to extract common aberrations caused by sourcing them online:

1. **Discard non-text documents** — this ensures that all documents are either plaintext or HTML;
2. **Convert to text** — the html2txt(Swartz (2010)) tool was used to render the HTML into a text-only form that approximated the original layout and document flow;
3. **Segment words** — each file was split into a bag of words for analysis;
4. **Filter by language** — only English language texts were analysed, classified as English if they contained five or more words from the Glasgow stoplist(taken from the TERRIER project, Ounis,I.;Lioma,C.;Macdonald,C.;Plachouras,V. (2007)).

At the end of this process, there were 132,000 documents in the corpus comprising 5.2 million terms. Attached to these documents were 1.2 million tags.

3.2 Tag Preparation

The Delicious.com tags required normalisation in order to make them comparable to terms from the documents. This processing of tags was relatively heavy, designed to correct a set of common user foibles such as incorrect use of separators when using the Delicious.com interface, or use of compound terms:

1. **Remove trailing commas** (resolves a common user error where users separate their tags like items in a list);
2. **Split the tag** using common delimiters, and apply CamelCase[1] detection;
3. **Check spelling of the tag**. If the tag fails a spellcheck using the GNU aspell US English dictionary:
   (a) List all cut-points where the tag is split into two English words (again according to aspell);
   (b) Select the one where both candidate words are of the most similar length;
4. **Convert to lowercase**;
5. **Process using Porter's stemming algorithm Porter (1997)**.

This processing resulted in a set of tag-terms, designed to be comparable to the native document terms. As an example, the most common tag, `webdesign`, would be split into `web` and `design`. For a more thorough example, see Figure 1. Together, the methods above account for around 97% of compound tag usage (Guy and Tonkin (2006)).

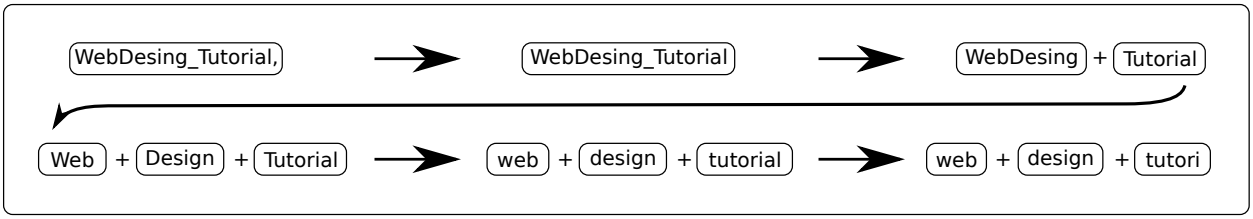


Figure 1: *The process of normalising tags.*

3.3 Metrics

Metrics for comparison were chosen to represent various approaches to measuring salience. Metrics such as `TFIDF` and `LOGLIK` are commonly used in the extraction phase of classifiers in order to identify significant topics, and it is these that are in theory the closest to tag measures. Two (`POFO`, `NDEGREE`) are included to represent specific features of the text which are otherwise omitted by the use of a bag-of-words model.

`NDEGREE` is the only metric based heavily on semantics. It is built by forming a graph between every noun and verb in a document, connecting words that share a synset in the WordNet (Fellbaum (1998)) dictionary, then taking the degree of each node in the graph. This forms a measure of the connectedness that spans polysemic words, offering a rough indication of topics.

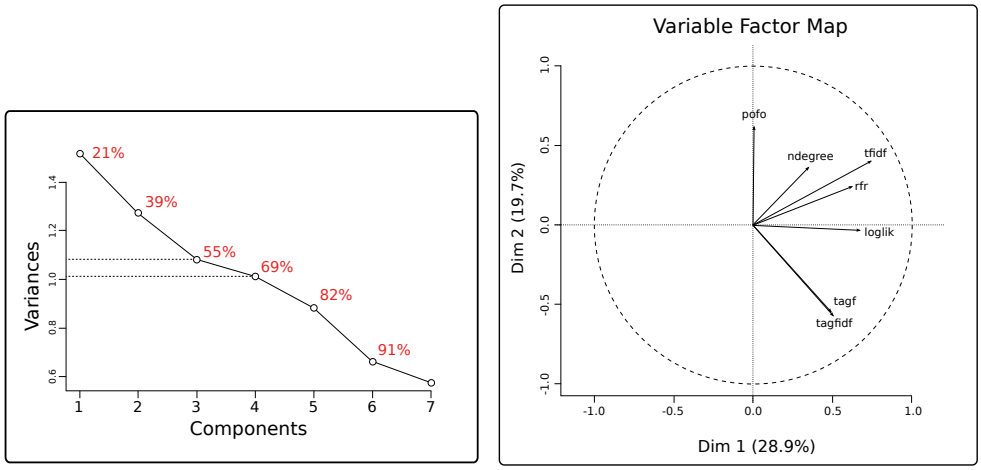
These term-rating methods are contrasted primarily against the tag metrics, which are computed by assuming that the documents 'contain' their tags (Hassan-Montero and Herrero-Solana (2006)). Computations are performed using statistics from the whole collection of documents, rather than each user's personal corpus. This more accurately represents the strategy chosen in conventional corpus linguistics, though arguably deviates from the normal use-case for document retrieval in the Delicious.com system.

Metric	Full Name	Description
`LOGLIK`	Log likelihood	As used in WMatrix (Rayson (2008)).
`TFIDF`	TF - IDF	A commonly used information retrieval measure.
`POFO`	Position of first occurrence	Used to weight titles and topic sentences.
`RFR`	Relative frequency ratio	The ratio of a term to the document length.
`NDEGREE`	WordNet Node degree	The node degree of a term in a graph built from WordNet synsets.
`TAGFIDF`	TFIDF using tag corpus	Designed to emulate TFIDF for tag measures.
`TAGF`	Tag frequency	A count of how many times a tag is used to describe a document.

Table 1: *The metrics used for comparison.*

3.4 Comparisons

In order to compare the metrics, they were processed into a series of tag and term lists, scoring each term relative to the original document within the corpus. These lists were analysed using



(a) *The PCA scree plot.*

(b) *First two components of the PCA per-metric.*

Figure 2: *The PCA model parameters.*

techniques that assert linearity, and so the logarithm of some metrics (`TFIDF`, `RFR`, `TAGFIDF`) was taken in order to foster this condition.

To explore and identify covariance between metrics, principal component analysis was performed upon the salience measures. This produced a set of equations, each expressing a component of variation within the text. The most active of these were chosen as a simplification of the model, and the general degree of agreement between each metric was used, along with their method of computation, to comment critically on this simplification.

Secondly, two sets of classifiers were built, each of which would select terms from the data set to be used as tags. One classifier was trained for each individual metric, as well as those based on the simplified all-metric model from the PCA findings. Crossvalidation was used with a sample of roughly 26.4 thousand documents to validate and compare the performance of these classifiers in a task that, roughly, equates to autotagging documents.

4 Exploratory PCA

Performing a princpal component analysis on all of the metrics produced the scree plot as shown in Figure 2a. This shows that there is relatively little colinearity between metrics, paying testament to the complexity of language and the pragmatic nature of salience measures used for keywording.

The factor map in Figure 2b displays the first two components identified by the PCA, illustrating a clear grouping between metrics that purport to measure similar quantities across these two components[2]. This is a rough indication that the fit, though complex, is still indicative of the underlying data, though the first two components explain only 40% of observed variance, which rises to 69% once common heuristics for selecting component limits (as stated in Duntemman (1989)) have been taken into account. In order to create a model that is likely to be useful, a four-factor solution was chosen. The loadings for each metric in this four-component model (after orthogonal rotation using the varimax algorithm) are listed in Table 2.

Metric	Comp.1	Comp.2	Comp.3	Comp.4
loglik				0.903
log.tfidf	-0.477		-0.433	0.239
pofo	0.119		-0.877	
log.rfr	-0.700		0.188	
log.tagfidf		-0.691		
tagf		-0.722		
ndegree	-0.509			-0.343

Table 2: *The chosen four-component solution from exploratory PCA.*

4.1 Discussion

From these loadings, it is clear that there is little in the way of simple structure: component 2 identifies the tag measures as quite separate from the others, as they lack any significant interactions with other components. `NDEGREE` is shown by component 1 as having greater association with the `TFIDF` and `RFR` measures of occurrence, despite its different philosophy. Component 4 is the only one involving `LOGLIK`, something alluded to by the biplot in Figure 2b. Interestingly, component 3 serves to illustrate the difference between `POFO` and the other measures, whilst suggesting similarity between it and `TFIDF`.

Identification of the principal components has offered a reasonable explanation of the two tag measures. They are shown to be quite different, varying in a dimension not observed on the other measures. This suggests that efforts at classifying will require metrics based on more complicated models of text than the ones analysed here, presumably one capable of topic-level determination.

Unfortunately, the assumptions made by maximum-likelihood based factor analysis make this four-component solution difficult to verify in practice due in part to its complexity, and in part due to the differences in distribution between metrics. We will, in using this model as the basis for a classifier in later sections, seek instead to identify its value through practical means.

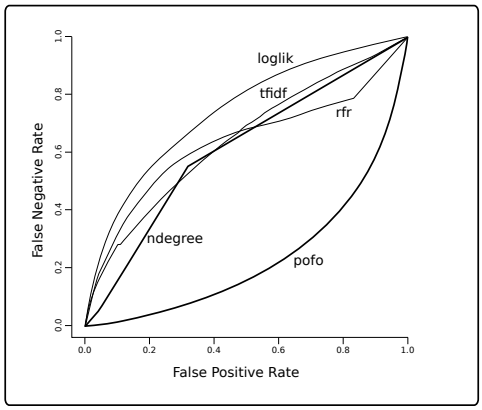
5 Tag Selection/Prediction

In order to assess the predictive capacity of each metric (as well as the exploratory PCA's model), each non-tag metric was taken as a predictor in its own right, simply using the proportion of its output relative to its maximum as the predicted probability that $tag\ frequency > 0$.

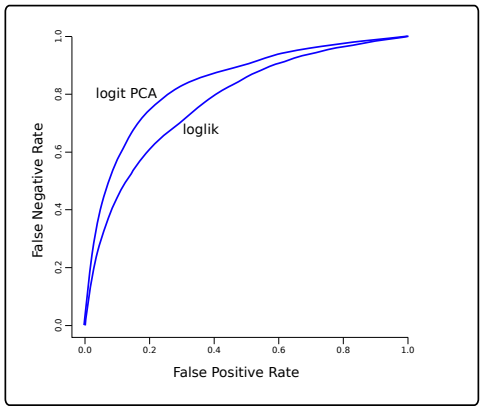
These classifiers for the individual metrics were taken as a baseline, to which was compared the model produced in the exploratory analysis. The PCA's four components were combined using a logistic regression model, again predicting that each term should be included at least once in the tag set.

5.1 Individual Metrics as Predictors

As can be seen from Figure 3a, `LOGLIK` is the best single predictor, producing a fairly smooth and reliable classification of tag use. The next best, exhibiting a similarly reliable shape due to its similar normalisation procedure, is `TFIDF`, followed by the other less reliable or less well normalised metrics.



(a) *ROC curves for each individual non-tag metric.*



(b) *ROC curves for logistic regression based on the exploratory PCA vs. raw* LOGLIK *scores.*

Figure 3: *Classifier ROC curves.*

5.2 Logistic Regression

Comparison of log likelihood to the PCA-based composite predictor illustrates just how good LOGLIK is at extracting pertinent keywords — the addition of the other components, designed to identify variation between other metrics, produces only a slight increase in the power of the predictor.

5.2.1 Discussion

Though the composite model is able to offer slight improvements over raw LOGLIK scores (as illustrated in Figure 3b), it is significantly more complex to apply in practice, requiring calculation of other metrics and their combination. The composite model achieves an $F_{0.5}$-score of 0.113, whereas the log likelihood metric alone manages 0.105.

A k-nearest-neighbour classifier was also fitted to the data (using $k = 3$) in an effort to discard the requirement for linear separability. This yielded very poor performance, roughly akin to that of the NDEGREE metric alone (straight ROC curves with AUCs around 0.55).

6 Conclusion

It is clear from the results of the classification that prediction of manually-assigned tags is possible to some degree. Since the predictive power of the composite model fails to exceed that of commonly used existing tag metrics by any large margin, it seems reasonable to propose that the variation identified by the model (and predicted) is largely descriptive of the document.

The ability to predict tags fully includes the ability to identify keywords which do not exclusively exist within the document (or even those similar to it). It is plausible that we have taken the first steps to identifying and classifying two major types of tag: *descriptive* tags, intended to identify the document for later retrieval, and *evaluative* tags, which involve some human intent or judgement and are most analogous to the category and higher-level tags from other taxonomy literature. This separation is indicated in part by our exploratory analysis, which creates a component for tag measures, into which no other metric significantly deviates.

Expanding the selection of metrics used to include models that make heavy use of semantic

and cultural clues in a similar way to discourse analysis may provide this power, though the degree to which information must be added, rather than selected from an existing corpus, is unclear.

It is clear that, for anyone attempting to predict manual tags, or to identify the degree to which they are coupled with their document, the only practical choice is log likelihood. Minor improvements may be achieved over the log likelihood metric by better identifying dimensions, though the value of this is may be outweighed by the increase in complexity.

6.1 Further Work

The possibility of being able to identify which tags most regard a document's contents leaves open the task of creating a more subjective tag generator using data extrinsic to the document. It is possible that, driven by the log likelihood selection of some tags, a hybrid tag predictor could be produced that would be cable of classifying documents in a similar manner to a human manually tagging them. This would open the door to topic-based browsing of large collections of data in a way that currently requires much manual intervention.

A comparison between the folksonomies of tags and search engine retrieval methods may also answer questions regarding how humans naturally organise their data. Adjusting the distance metrics used for searching to more accurately portray an individual's opinion of topic-space would lead to a more natural search interface, something Google and others are currently working towards with their personalised search options.

It is possible that much of the ambiguity remaining in tag sets may be extracted through the use of more complex linguistic models. Many of the metrics used in this study focused on the Bag-of-Words model, which fails to adequately represent larger-scale features of the document, such as semantic field or purpose.

Notes

[1]A way of writing phrases whilst omitting spaces, instead capitalising each word.
[2]The map was computed on a subsample of 100,000 terms due to the low speed of rendering.

References

Dunteman, G. "Principal component analysis. quantitative applications in the social sciences series (vol. 69)", 1989.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

Guy, M. and E. Tonkin. (2006). "Folksonomies: Tidying up tags?". *D-lib Magazine*, 12(1).

Hassan-Montero, Y. and V. Herrero-Solana. (2006). "Improving tag-clouds as visual information retrieval interfaces". In *International Conference on Multidisciplinary Information Sciences and Technologies*, 25–28. Citeseer.

Kehoe, A. and M. Gee. (2011). "Social tagging: A new perspective on textual aboutness".

Mathes, A. (2004). "Folksonomies-cooperative classification and communication through shared metadata". *Computer Mediated Communication*.

Ounis,I.;Lioma,C.;Macdonald,C.;Plachouras,V. (2007). "Research directions in terrier". *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*.

Porter, M. F. (1997). "An algorithm for suffix stripping". 313–316.

Rayson, P. (2008). "From key words to key semantic domains". *International Journal of Corpus Linguistics*, 13(4), 519–549.

Scott, M. (1996). "Wordsmith tools". *Software for Windows*, 3, 95.

Swartz, A. "html2text", 2010. `http://www.aaronsw.com/2002/html2text/`.